6) Encoder–Decoder with Masked Attention

a) Why Masked Self-Attention in the Decoder?

The **decoder uses masked self-attention** to ensure that each position can only attend to **previous tokens** (and not future ones).
This prevents the model from "cheating" by looking ahead at future words during training, thereby preserving the **autoregressive (left-to-right)** nature of text generation.

b) Encoder Self-Attention vs. Encoder–Decoder Cross-Attention:

Type	Description	Purpose
Encoder Self-Attention	Each token in the input sequence attends to **all other input tokens**.	Builds a contextualized representation of the entire input sequence.
Encoder–Decoder Cross-Attention	The decoder's queries attend to the **encoder's output representations (keys and values)**.	Allows the decoder to focus on relevant parts of the input while generating output tokens.

c) Inference (No Teacher Forcing) – Step-by-Step Generation:

During inference, the decoder starts with a special `<START>` token and generates **one token at a time**.
After each prediction, the new token is **fed back into the decoder** as input for the next step, continuing until an `<END>` token is produced or a maximum length is reached.
This process is **auto-regressive**, meaning each new token depends on all previously generated tokens.